HLT for HL-LHC, Technology and architecture for next decade TDAQ

Silvia Amerio

University of Padova and INFN

IFD2014 – INFN workshop on Future Detectors for HL-LHC
Trento 11-13 March 2014

Introduction

High Level Triggers in LHC experiment are based on *farms of CPUs*

- easier maintenance and upgrade
- the HLT code can be easily simulated offline
- HLT code as much similar to the offline as possible
- additional computing power for offline processing when the experiment is not taking data

New challenges in the future:

- LHC upgrade → Increased event rates and pile-up → more computing power needed
- We are facing a technological change: many-cores, parallel computing.

In this talk:

- *Alice and LHCb current HLT systems and plans for Phase 1 upgrade*
  - *Towards a triggerless event selection*
  - *Usage of latest technologies for data movement and data processing.*
- *Prospects for Atlas and CMS in Phase 2*
- *INFN groups involved*

ALICE upgrade strategy

Requirements

Physics objective of the Phase 1 upgrade is the ***precision measurements*** of the QGP, which will be accessible through measurements of heavy-flavour and quarkonia ***down to zero pT and low mass di-leptons***.

Cannot be triggered at hardware level --> *data taking with minimum-bias trigger or continuous mode*

Increased statistics (at least 10 nb^{-1}, 100 times current sample) –> *Sample full 50 kHz PbPb interaction rate (9.2 Tbit/s detector readout)*

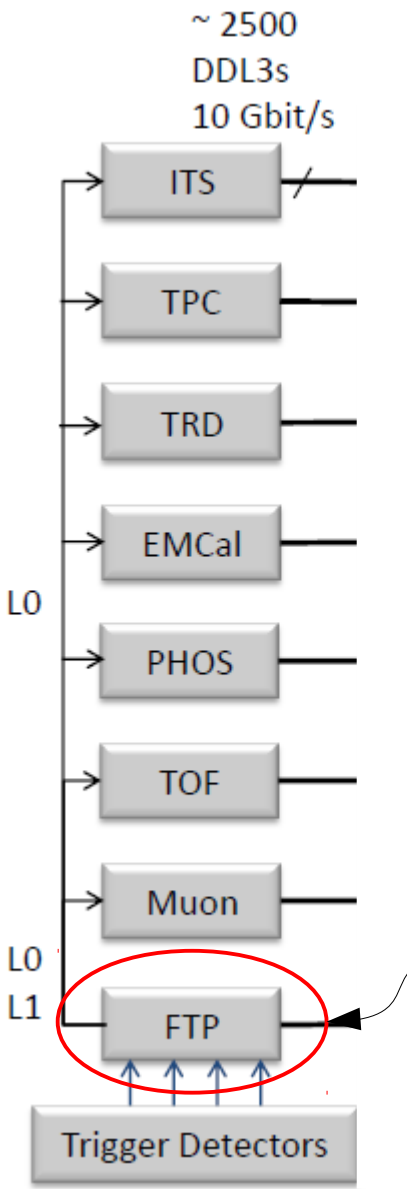
Strategy

Data reduction by (partial) online reconstruction and compression

Store only reconstruction results, discard raw data
- Demonstrated with TPC clustering since Pb-Pb 2011
- Optimized data structures for lossless compression
- Algorithms designed to allow for offline reconstruction passes with improved calibrations

Implies much tighter coupling between online and offline reconstruction software

Upgrade architecture

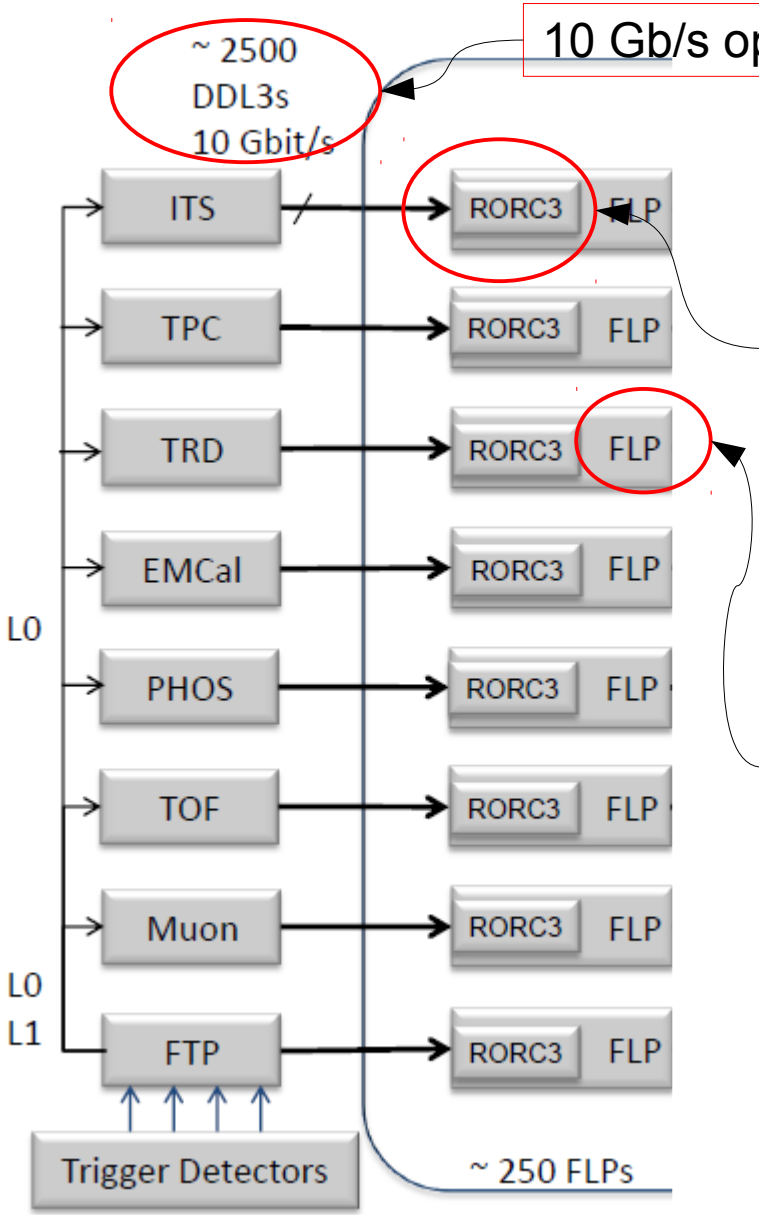


DETECTOR READOUT

Continuous readout for the Time Projection Chamber (TPC) and the Inner Tracking System (ITS) (at 50 kHz, ~ 5 events during drift time of 92 μs)
Triggered readout for all other sub-detectors

FTP (Fast Trigger Processor) : it provides clock/L0/L1 to triggered detectors and TPC/ITS for data tagging and test purposes.

Upgrade architecture



10 Gb/s optical links

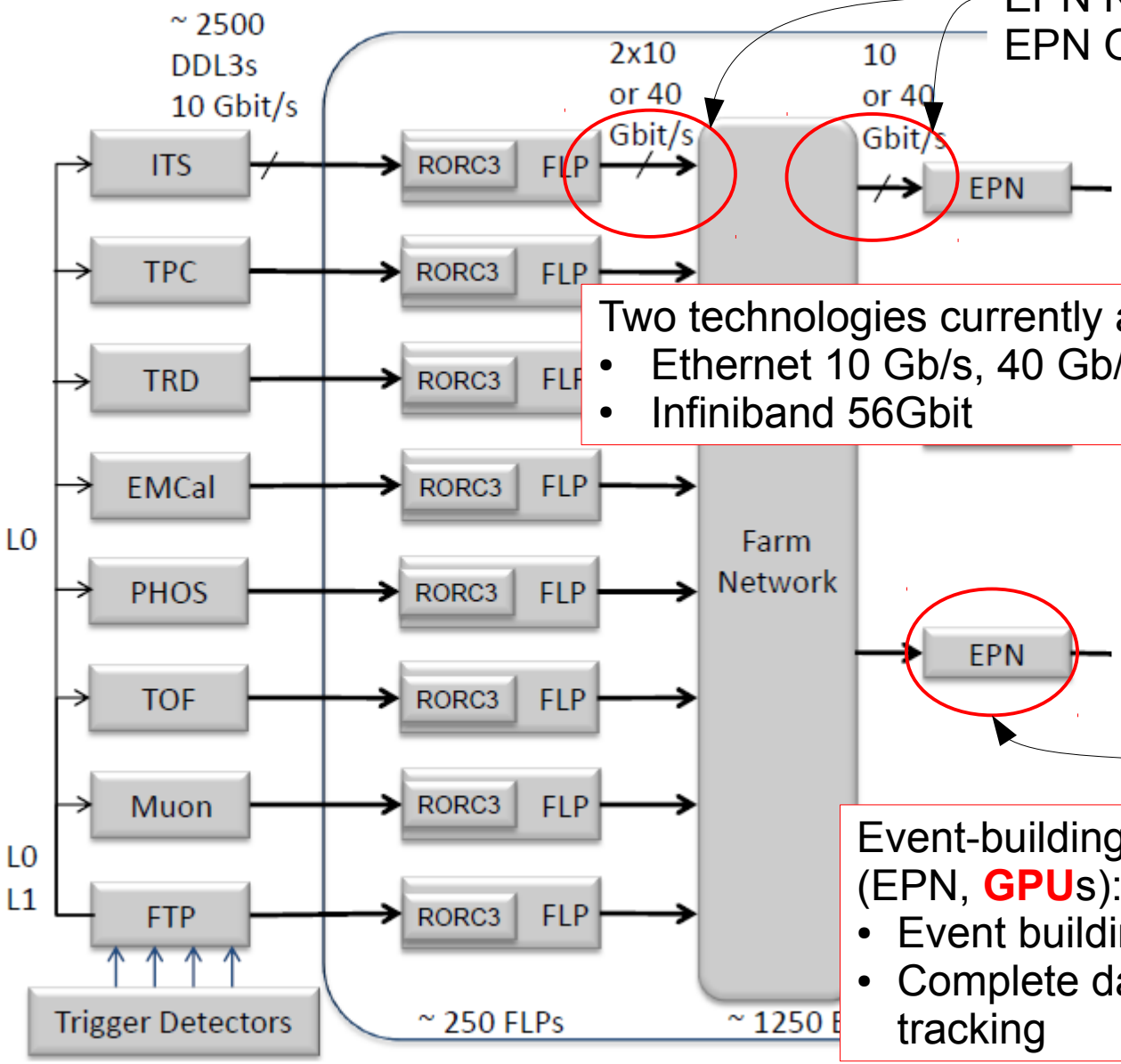
On the fly data compression in the Read-out receiver cards (RORC3, PCIe Gen 3)

First Level Processor (FLP, CPUs):

- First step of data processing (clusterization and local tracking)

Upgrade architecture

FLP Node Output: up to 12 Gbit/s
EPN Node Input: up to 7.2 Gbit/s
EPN Output: up to 0.5 Gbit/s

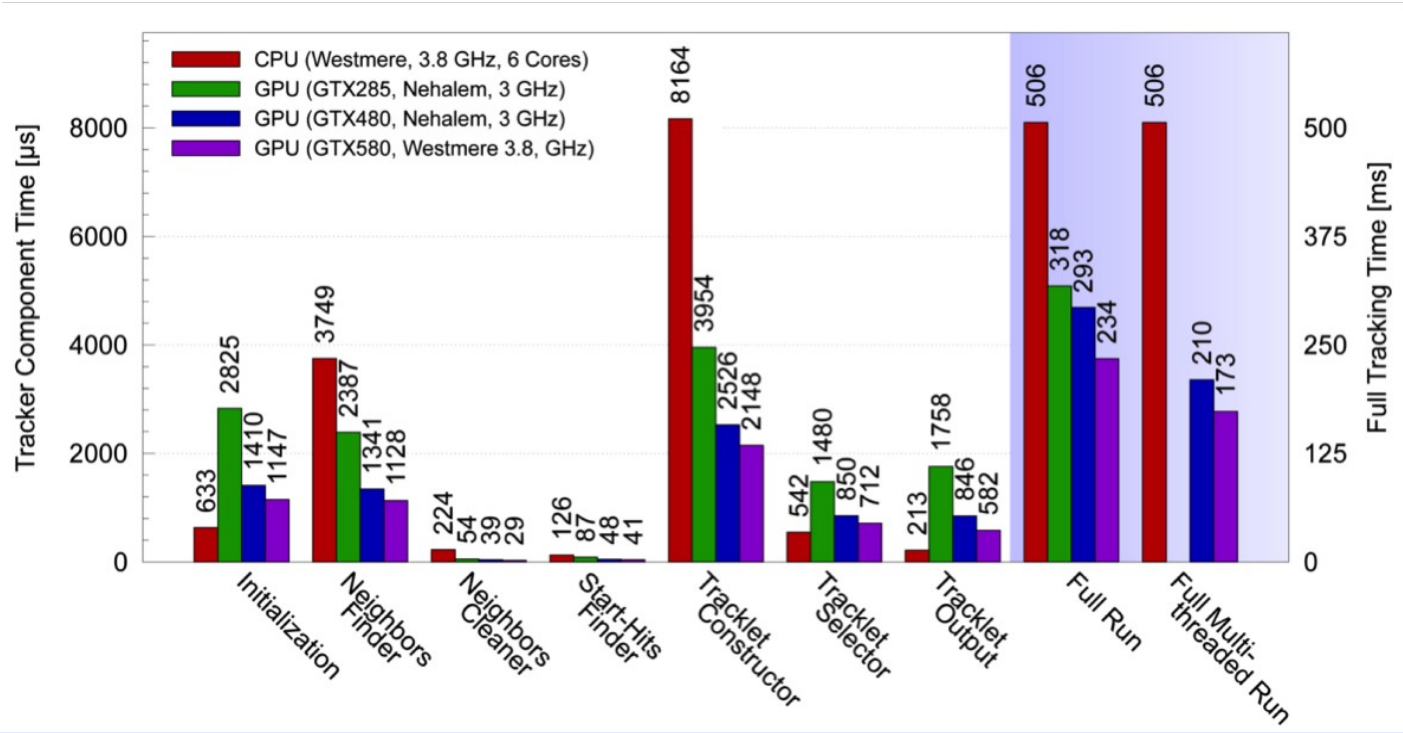


Two technologies currently available
- Ethernet 10 Gb/s, 40 Gb/s
- Infiniband 56Gbit

Event-building and Processing Nodes (EPN, **GPU**s):
- Event building
- Complete data processing, include global tracking

GPUs in Alice HLT

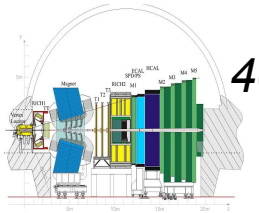
TPC tracking most time consuming step in the HLT
- Neighbor Finding, Tracklet Construction and Tracklet Selection on GPU;
- Inizialization, Tracklet Output on CPU.
- *Overall total processing time from 500 ms to 170 ms.*



Lessons learned:
- Best results on a *hybrid system (CPU/GPU)*
- CPU and GPU versions can run both online and offline → *the consistency of results between the two versions can be easily checked*

LHCb upgrade: motivations

No deviations from the SM seen yet, we need to ***probe NP at scales >> 1 TeV*** → **increase the precision** of the measurements → ***increase the size*** of the collected signal samples → ***LHCb upgrade to collect up to 50 fb^{-1}***



40 MHz

L0
(custom electronics)

μ, e ,γ and hadrons p_T cuts

40 MHz

DAQ
(custom electronics

1 MHz

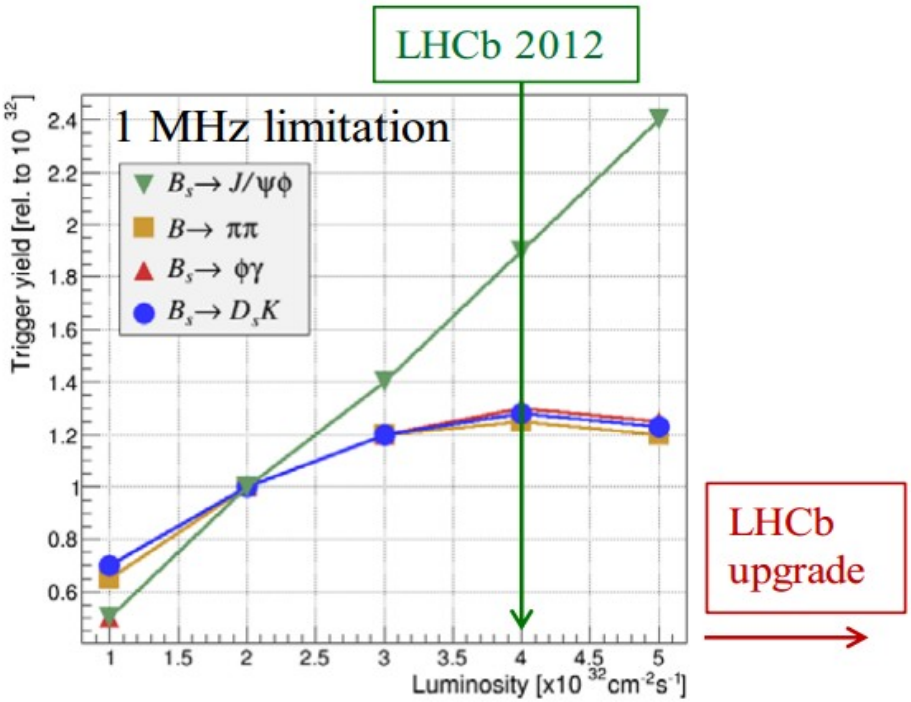
Event building
(CPU farm)

1 MHz

HLT
(CPU farm)

5 kHz

Storage



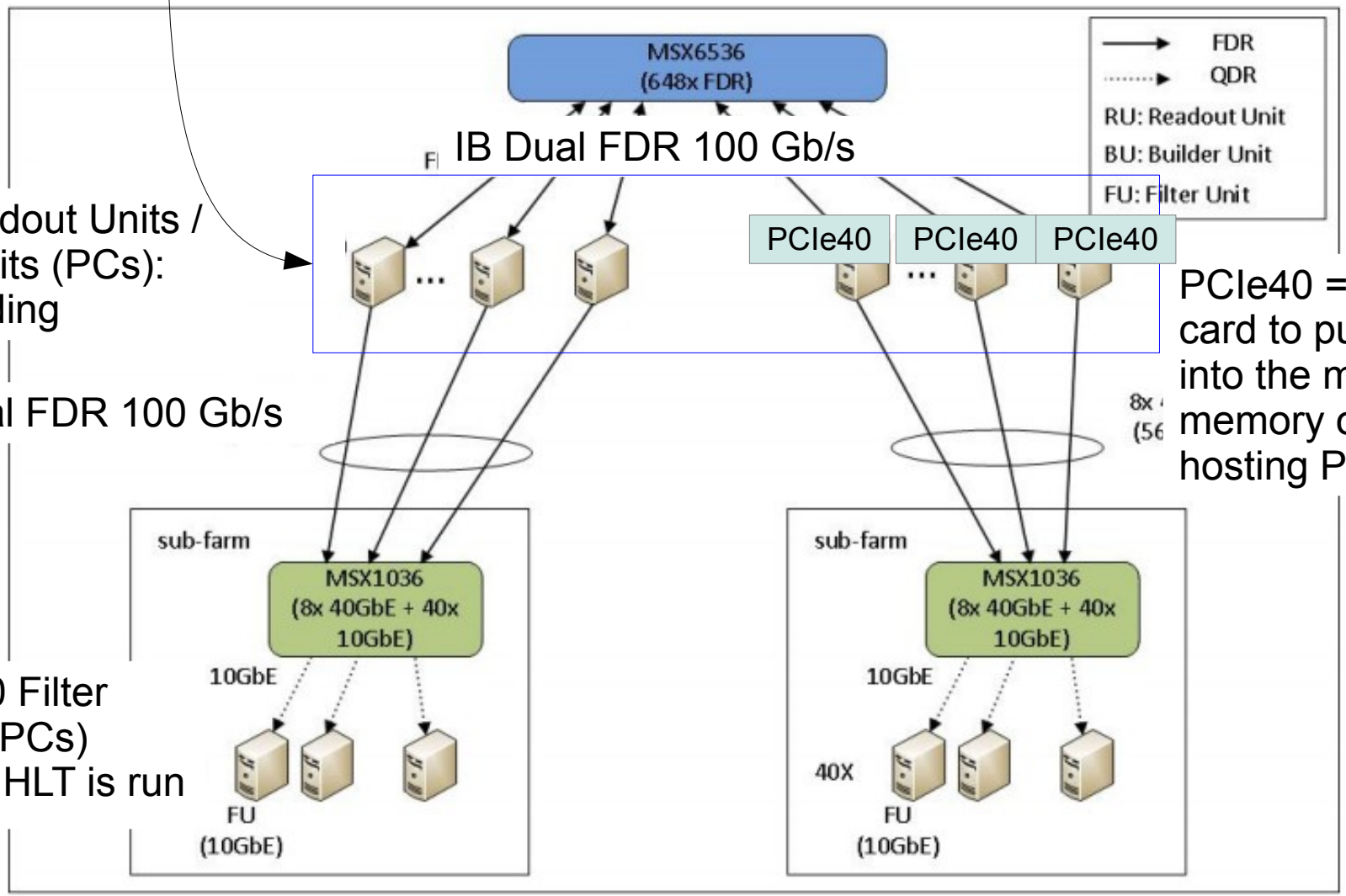
LHCb is currently limited by trigger and DAQ. Upgrade plan:

- Remove the L0 bottleneck → read out full detector at 40 MHz
- New DAQ architecture
- Low Level Trigger (LLT) for rate control out of DAQ
- *New EB/HLT*

New DAQ architecture

NB: Different architectures under study. This is the one currently considered the most cost effective.

100 kB x 40 MHz = 32 Tbit/s
to the DAQ via GBT links



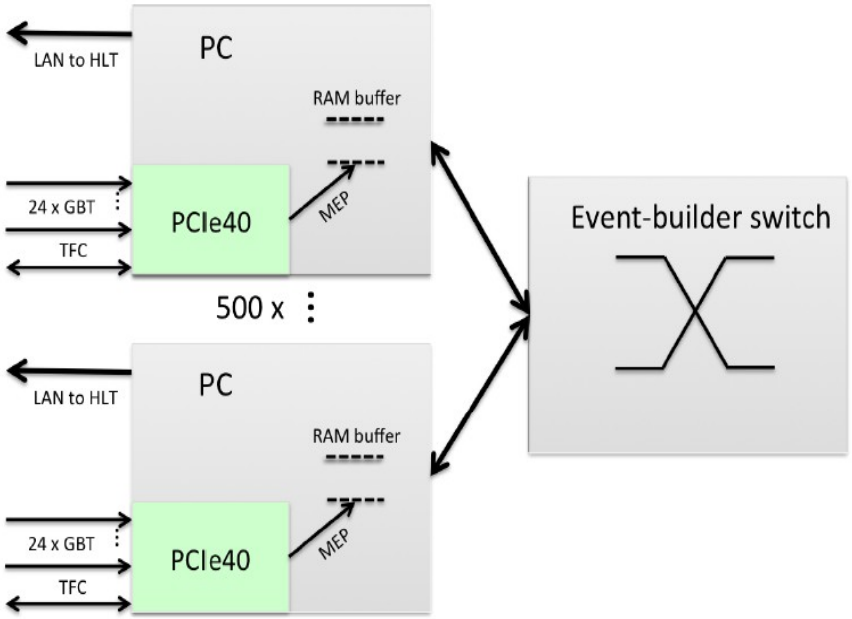
~ 500 Readout Units / Builder Units (PCs): event building

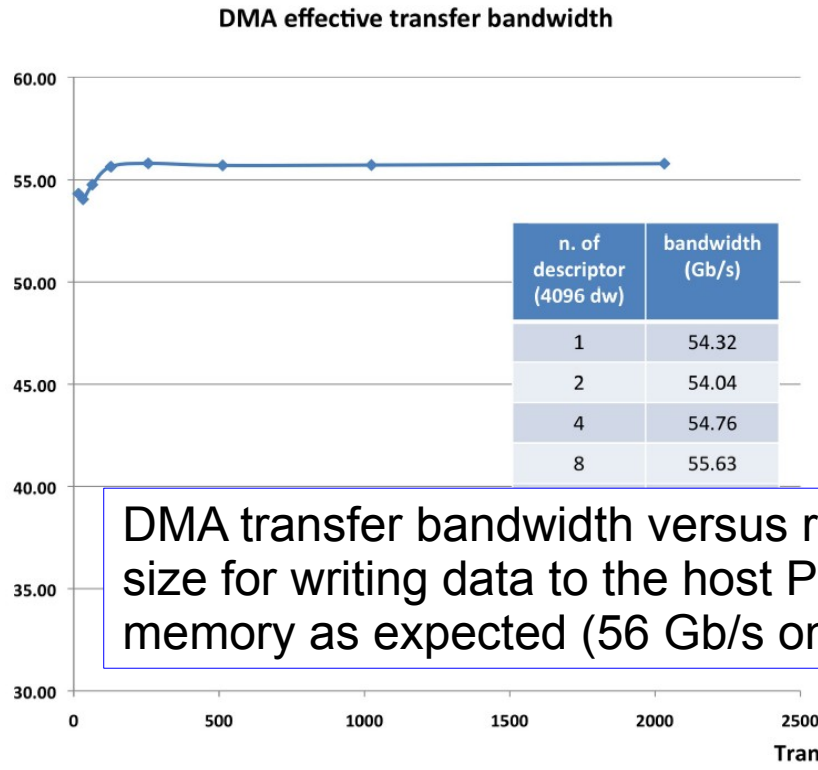
IB Dual FDR 100 Gb/s

~ 1000 Filter Units (PCs) where HLT is run

PCIe40 = PCIe card to push data into the main memory of the hosting PC

PCIe based RO and EB

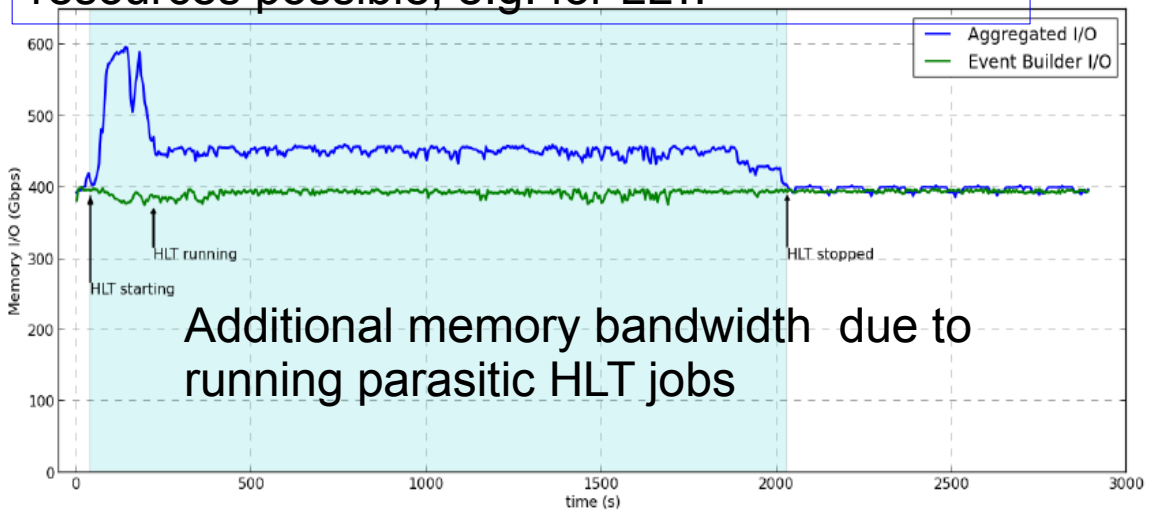




n. of descriptor (4096 dw)	bandwidth (Gb/s)
1	54.32
2	54.04
4	54.76
8	55.63

DMA transfer bandwidth versus record size for writing data to the host PC memory as expected (56 Gb/s on 8x)

EB performs stable at 100 Gb/s on one node (total throughput 400 Gb/s on a single server)
Parasitic usage of the event builder units resources possible, e.g. for LLT.



Additional memory bandwidth due to running parasitic HLT jobs

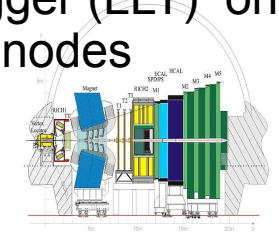
Advantages of PCIe:

- *It is the interconnect choice for many computing and network technologies (GPU cards, IB or 40Gb Ethernet, solid state disks,..)*
- *4th generation under development - backward compatible.*

Possible strategies for the EB/HLT

A

Same as as in the LOI, but with a Low Level Trigger (LLT) on the EB nodes



40 MHz

DAQ (custom electronics)

40 MHz

LLT on EB FPGA or EB CPU
Track pre-processing on FPGA (see Annovi's talk)
Event building on CPU

1-40 MHz

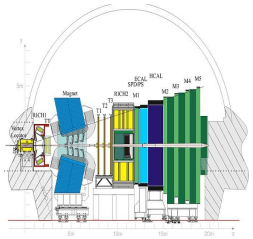
HLT (CPU only)

20 kHz

Storage

B

With GPUs in the EB nodes



40 MHz

DAQ (custom electronics)

40 MHz

Pre-processing with GPUs in EB nodes.
Event building on CPU

1-40 MHz

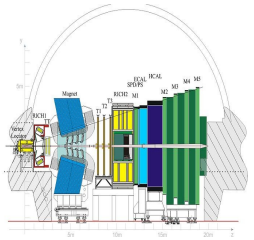
HLT (CPU only)

20 kHz

Storage

C

With GPUs both in the EB and HLT nodes



40 MHz

DAQ (custom electronics)

40 MHz

Pre-processing with GPUs in EB nodes.
Event building on CPU

1-40 MHz

HLT with CPU/GPU, e.g. to speed up particle identification algorithms

20 kHz

Storage

Scenario A: Improving HLT code

Requirements:

- Increase efficiency on hadronic modes

mode	D → hhh	B→ hh
ε(L0) [%]	27	62
ε(HLT/L0) [%]	42	85
ε(tot) [%]	11	52

- After the upgrade background rejection is not the main problem anymore, 25% event will contain a b or c quark --> *we need to better categorize signal while reducing the background*
- Assuming a CPU farm 10x the current one the time budget for HLT ~ 15 ms/event

Fundamental ingredients of the upgraded HLT:

- *full forward tracking* as early as possible
- *offline-like particle identification*
- *retuning of inclusive and exclusive selections*

	Timing [ms]		
Tracking Algorithm	No GEC	GEC = 1000	GEC = 600
VELO tracking	2.2	1.7	1.0
PV finding	0.41	0.36	0.25
VELO-UT tracking	2.5	2.1	1.5
Forward tracking	4.3	2.6	1.3
Total	9.4	6.8	4.1
MinBias inefficiency	0	(10.1 ± 0.9) %	(41.1 ± 0.8) %

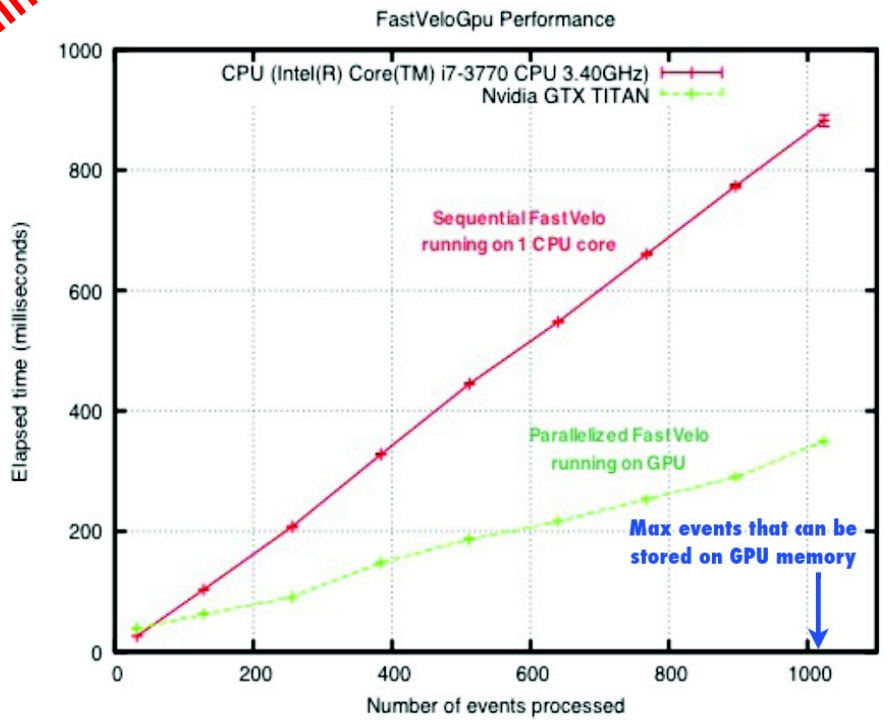
Scenarios B and C: investigating new technologies

Current VELO (Vertex Locator detector) tracking algorithm has been ported on CUDA and run on NVIDIA GPUs:

HLT1 Only	FastVelo on GPU		FastVelo	
	Efficiency	Clones	Efficiency	Clones
VELO, all long	89.2%	4.1%	89.6%	0.4%
VELO, all long, p > 5 GeV	92.0%	4.1%	92.1%	0.3%
VELO, all long B daughters	91.5%	4.0%	91.6%	0.5%
VELO, all long B daughters, p > 5 GeV	93.4%	4.2%	93.8%	0.5%
VELO, ghosts	25.8%		5.5%	

Efficiencies comparable to current algorithm; more ghosts, investigating...

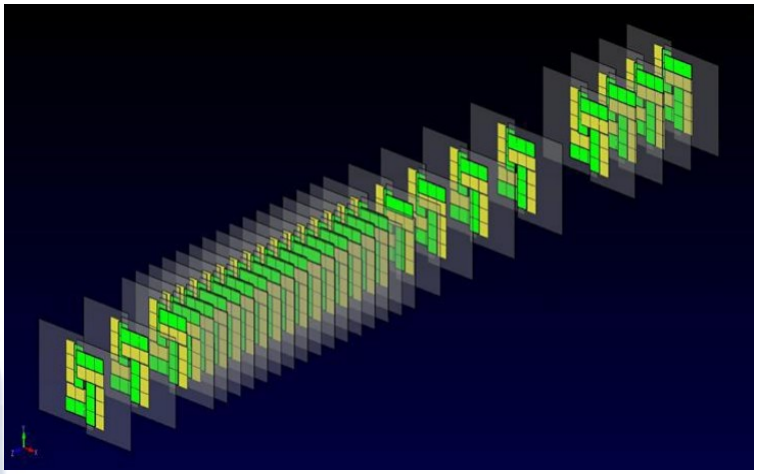
Preliminary

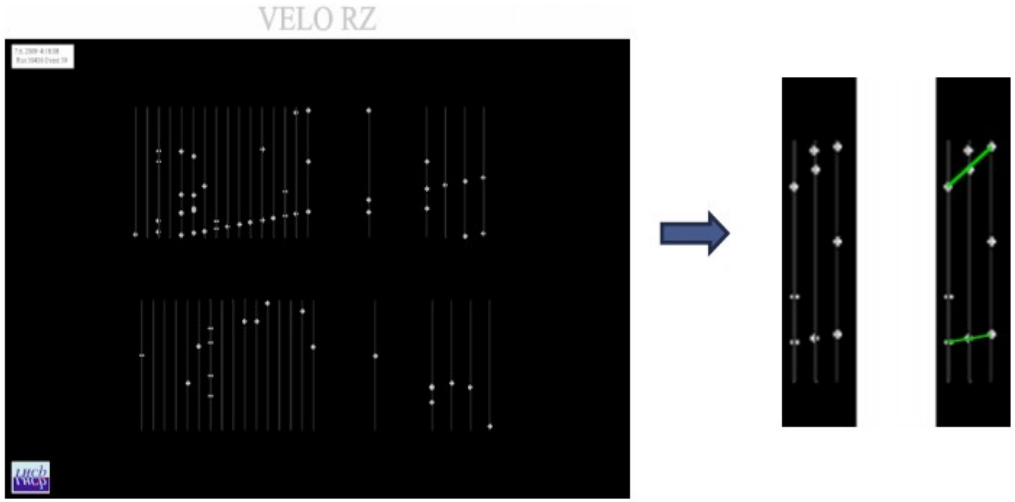


First results on timing are encouraging.

Goal: tests in parasitic configuration during Run 2

Scenarios B and C: investigating new technologies





Pixel (Upgrade) VELO Forward Tracking local search method on the GPU,

Triplets of clusters in neighbouring sensors are selected as seeding, forwarding is done in parallel.

Algorithm gets a reasonable speedup (11x), tracking performance indicators (ie. Reconstruction Efficiency, Clone and Ghost Fraction) need to be polished.

Preliminary

HLT in CMS and ATLAS for Phase 2

For phase 2 HLT will have to provide a rejection factor
- ATLAS: from 200 kHz to 5-10 kHz
- CMS: from 500-1000 kHz to 10 kHz

CMS estimates x50 increase in the required HLT processing power for phase 2.
Moore's law can lower this estimate by a factor 10.

Where can the missing factor 5 can come from?
- Code/framework improvement
- New technologies

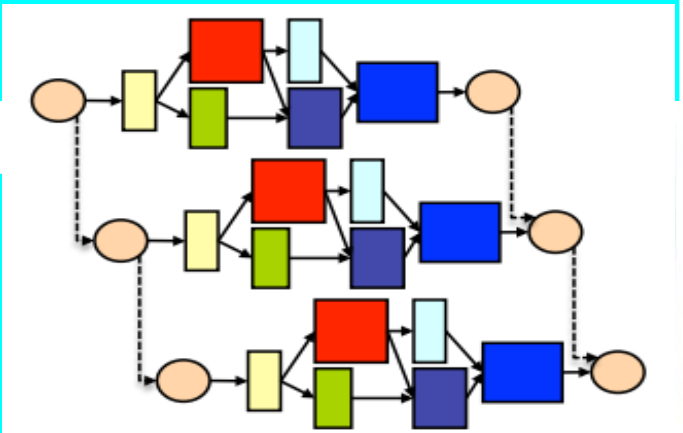
HLT in CMS and ATLAS for Phase 2

For phase 2 HLT will have to provide a rejection factor
- ATLAS: from 200 kHz to 5-10 kHz
- CMS: from 500-1000 kHz to 10 kHz

CMS estimates x50 increase in the required HLT processing power for phase 2.
Moore's law can lower this estimate by a factor 10.

Where can the missing factor 5 can come from?
- **Code/framework improvements**
- New technologies

Parallelism at increasing levels:
- ***Event level parallelism:***
  - 1 core – 1 event, fork + common memory shared using Copy-on-Write to reduce memory footprint
- ***Algorithm level parallelism :*** multi-threading
- ***Sub-algorithm level parallelism***



Thread-safe code
New frameworks, e.g.
- AthenaMP, GaudiMP, for event-level multi-processing
- New CMSSW, AthenaHive, GaudiHive, ALFA, for multi-threading

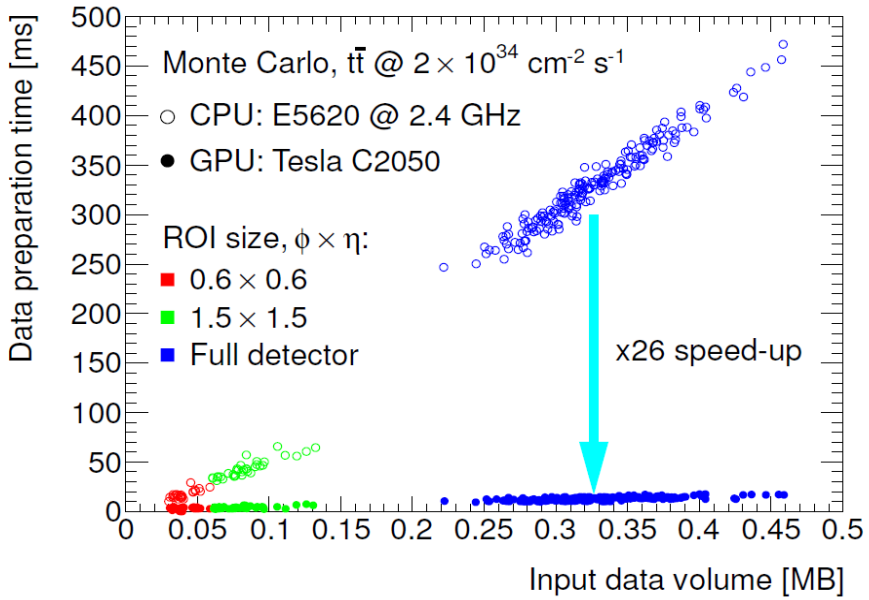
HLT in CMS and ATLAS for Phase 2

For phase 2 HLT will have to provide a rejection factor
- ATLAS: from 200 kHz to 5-10 kHz
- CMS: from 500-1000 kHz to 10 kHz

CMS estimates x50 increase in the required HLT processing power for phase 2.
Moore's law can lower this estimate by a factor 10.

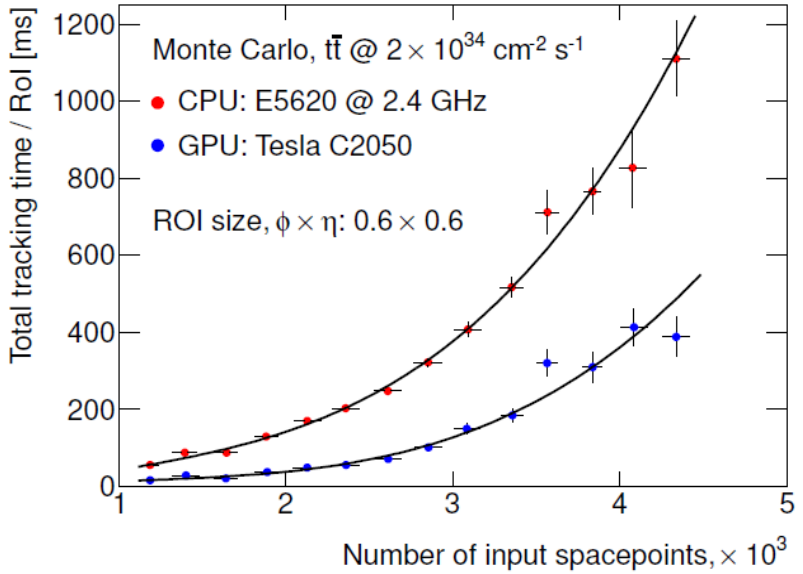
Where can the missing factor 5 can come from?
- Code/framework improvements
- **New technologies**

ATLAS HLT Inner Detector data preparation and tracking algorithms ported to GPU



Performance improvement for **data preparation steps.**

http://cds.cern.ch/record/1449116?ln=it



Performance improvement for **tracking steps**

HLT in CMS and ATLAS for Phase 2

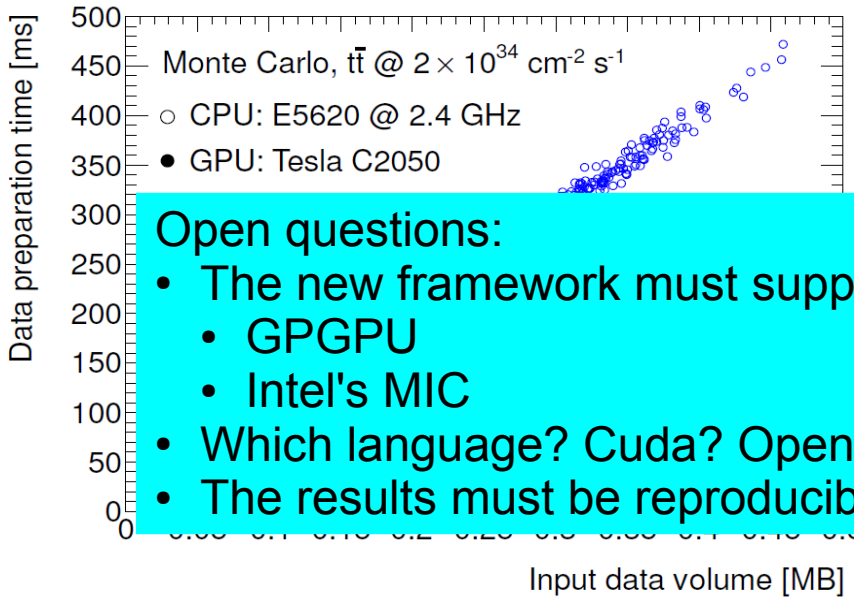
For phase 2 HLT will have to provide a rejection factor
- ATLAS: from 200 kHz to 5-10 kHz
- CMS: from 500-1000 kHz to 10 kHz

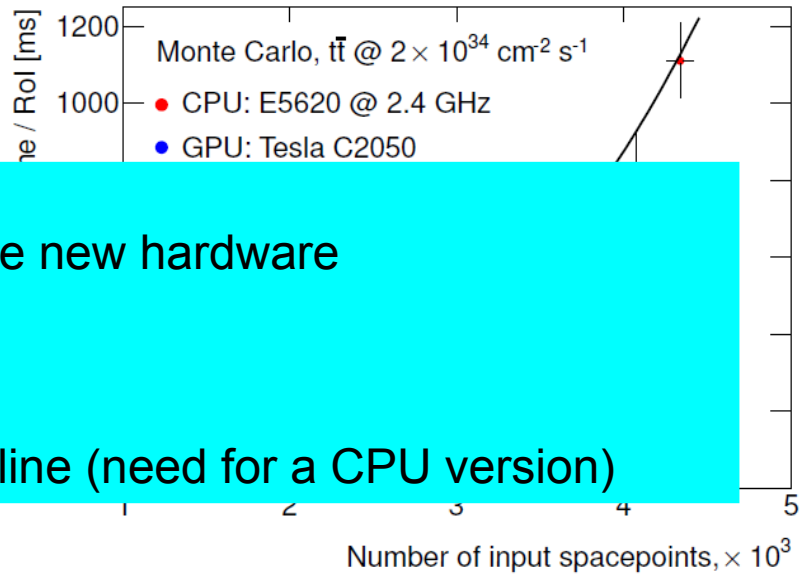
CMS estimates x50 increase in the required HLT processing power for phase 2.
Moore's law can lower this estimate by a factor 10.

Where can the missing factor 5 can come from?
- Code improvements
- **New technologies**

ATLAS HLT Inner Detector data preparation and tracking algorithms ported to GPU

http://cds.cern.ch/record/1449116?ln=it





Open questions:
- The new framework must support the new hardware
  - GPGPU
  - Intel's MIC
- Which language? Cuda? OpenCL?
- The results must be reproducible offline (need for a CPU version)

Performance improvement for **data preparation steps.**

Performance improvement for **tracking steps**

INFN groups involved in R&D activities

LHCb
- Bologna: DAQ readout upgrade based on PCIe Gen3
- Padova: HLT based on many-core technologies
- Pisa/Milano: Track pre-processing on FPGA

ATLAS
- Roma: Application of GPUs to ATLAS high level trigger (Interface between Athena and GPU, parallelization of L2-Muons algorithms)
- Bologna: Application of GPU/GPU-ARM to online tracking (Inner detector ITK)

Possible collaborations
- **CNAF** is willing to start an official R&D activity on new architectures; all LHC experiments could benefit in terms of resources and expertise
- In the past **LNL** contributed to CMS DAQ; any interest for the investigation of new technologies for CMS trigger?

Ongoing and future projects

GPU for real-time applications in HEP and medical imaging (GAP)
Call: FIRB 2012
Units: Pisa (PI), Ferrara, Roma

Big Physics Data on Arm and GPGPUs (BAG)
Call: H2020-MSCA-ITN-2014
Beneficiaries: Cern, University & INFN (PD/BO), ESRF, Funitec, Boston Ltd
Partners: ARM, Nvidia

Summary

High Level Triggers are facing big challenges in view of HL-LHC

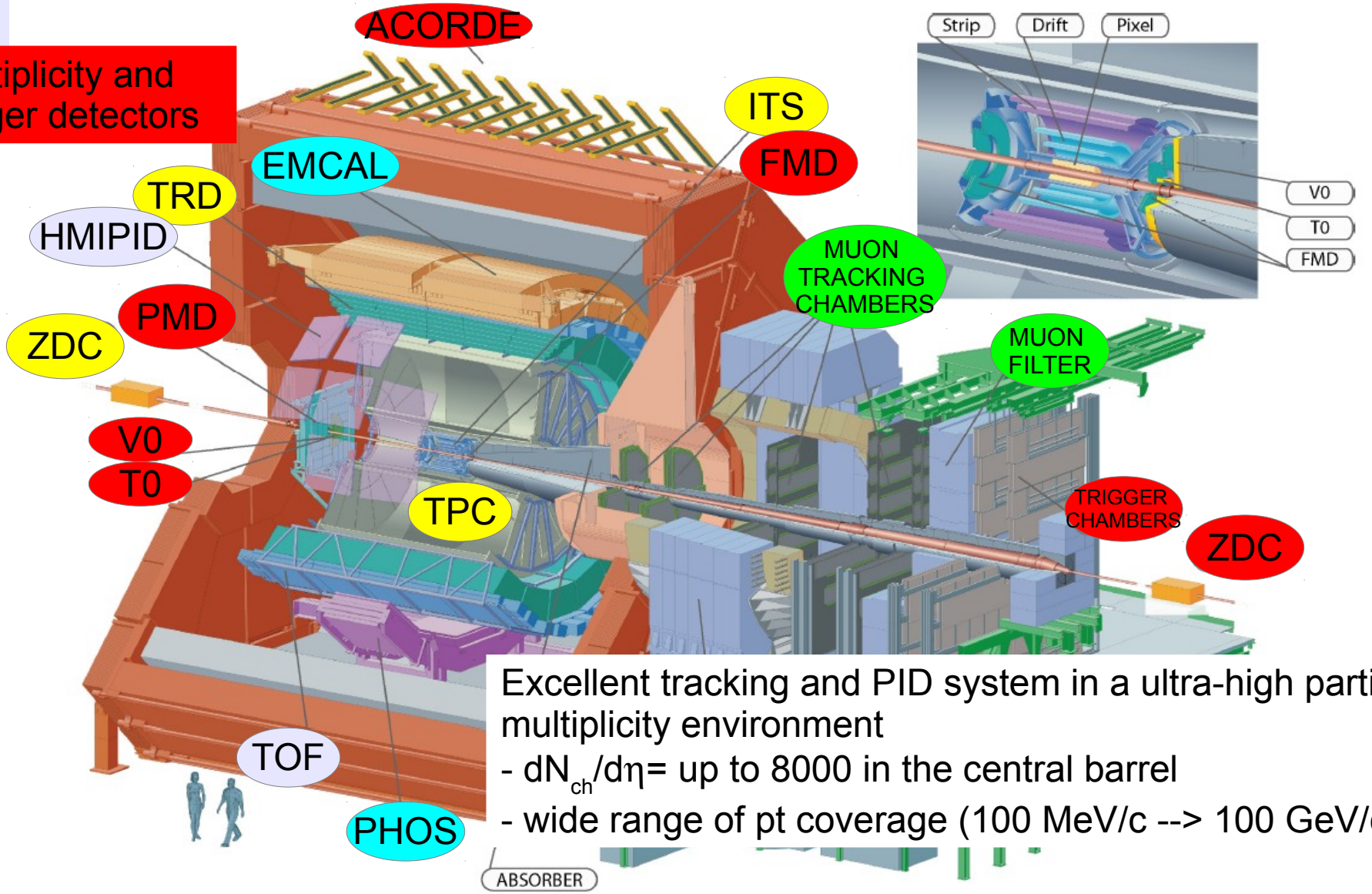
- Much higher rates (at higher pile-up)
  - *Alice and LHCb will deal with these problems already during Phase1. Current HLT upgrade designs are based on high-speed commercial networks coupled to an efficient usage of many-core technologies.*
- Adapt to new technologies
  - *Flexible DAQ architectures based on PCIe, to allow for a late decision on network and computing devices (IB/Ethernet, CPU/GPU/MIC,...)*
  - *Algorithms and frameworks need to be re-designed to allow for parallelism at all levels --> this applies to offline as well, big change in HEP software model.*

Interest in many INFN groups (not only HEP) for the application of new technologies to the online selection of events --> room for collaboration and sharing of experience.

BACKUP

Current ALICE detector

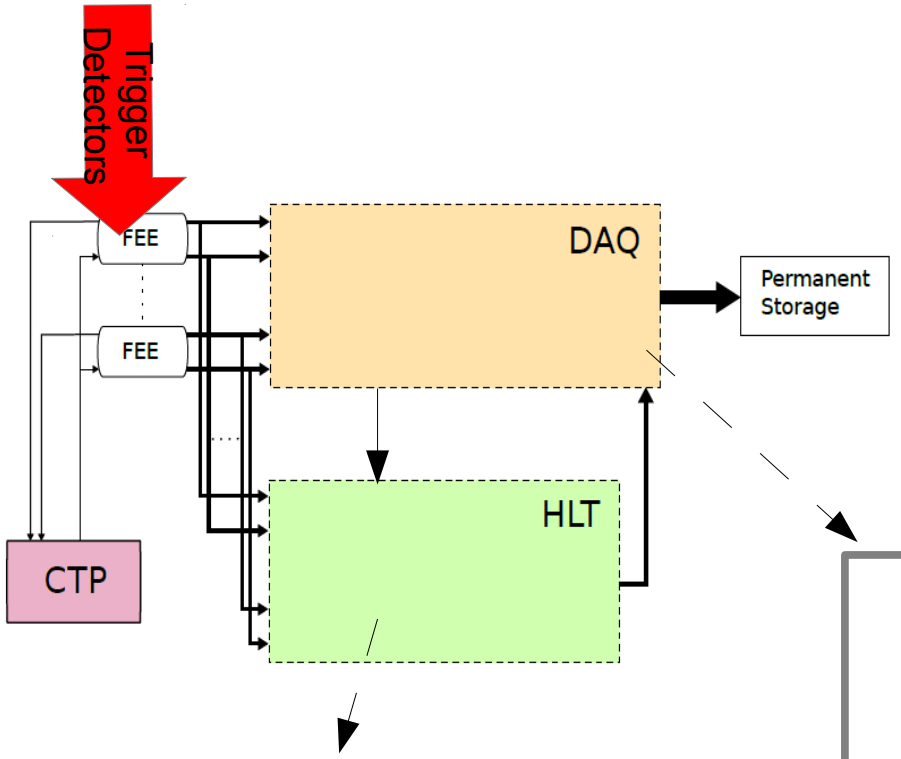
Tracking detectors
Calorimeters
Muon detectors
PID
Multiplicity and trigger detectors



Excellent tracking and PID system in a ultra-high particle multiplicity environment

- $dN_{ch}/d\eta$= up to 8000 in the central barrel
- wide range of pt coverage (100 MeV/c --> 100 GeV/c)

ALICE trigger system



- The Central Trigger Processor (CTP) collects trigger inputs from the trigger detectors, computes a decision and sends the read-out signal to all detectors.
- Upon a trigger by CTP, Front-End Electronics (FEE) sends data to *both the DAQ and HLT* --> symmetric interface to the detector readout

PC farm (2500 cores/200 nodes + 64 GPUs) which allows for complex trigger based on the full event information obtained from the reconstruction

Local Data Concentrators (LDCs, commodity PCs) for *sub-event fragments* building

Global Data Concentrators (GDCs, commodity PCs) for *complete event* building

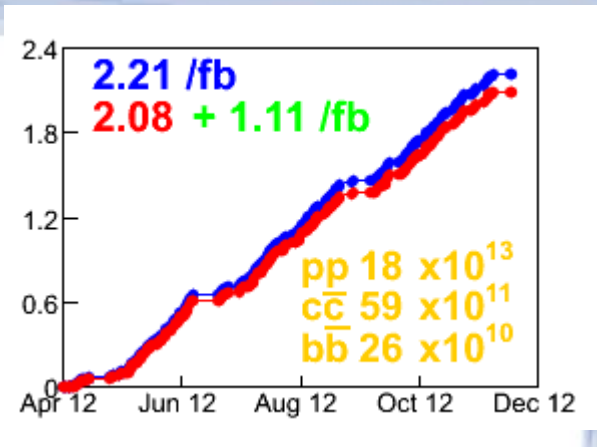
Different run modes (DAQ only, DAQ + HLT w/ and w/o HLT decision) reflecting an increasing usage of the HLT from commissioning up to the operation periods.

GPUs in Alice HLT

Lessons for the upgrade:

- Not all track reconstruction steps are well suited for GPUs --> *best results with a hybrid system, taking advantage from different technologies*
- The main tracking algorithm is contained in a common source code for both CPU and GPU versions; the code is included in specialized wrappers for each architecture and processed by the CPU and GPU compiler --> *changes to the code are minimized; it is easy to switch from one version to the other*
- The general tracker interface allows to use both GPU and CPU tracker within different frameworks: HLT and AliRoot (Alice offline framework for data reconstruction, analysis and simulation) --> *the consistency of results between the two versions can be easily checked; online system simulation can run on standard CPUs*

LHCb detector

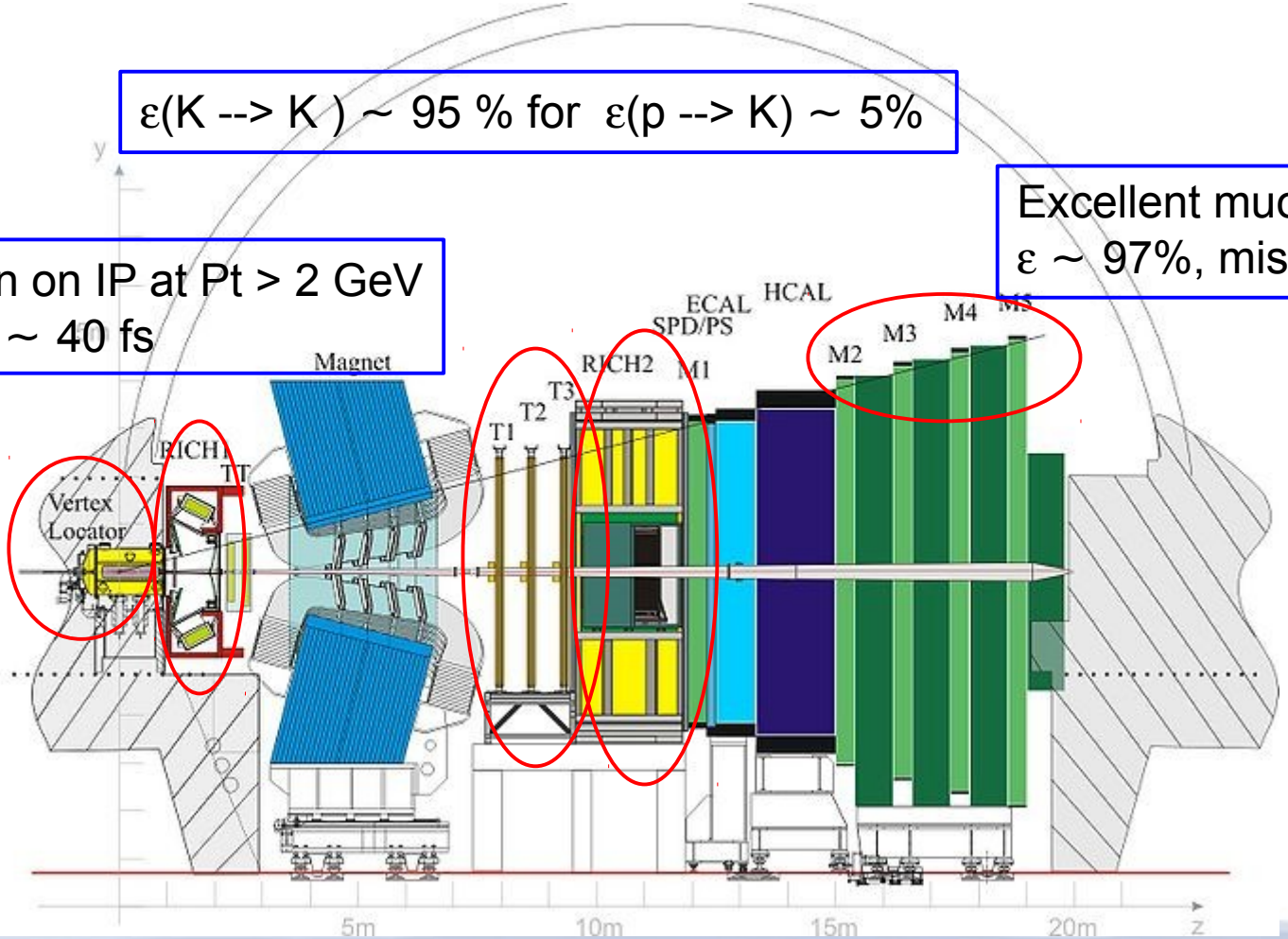


Designed to maximize the yield of B hadrons
Single arm spectrometer covering 2 < eta < 5
pp -> bbX ~ 75 μbarn @ 7 TeV, all types of B hadrons produced
cc-bar cross section = 20 x bb-bar --> also a charm factory.

ε(K --> K) ~ 95 % for ε(p --> K) ~ 5%

Excellent muon identification
ε ~ 97%, mis-id ~ 2 %

20 μm resolution on IP at Pt > 2 GeV
Time resolution ~ 40 fs



LHCb: data-flow in the EB

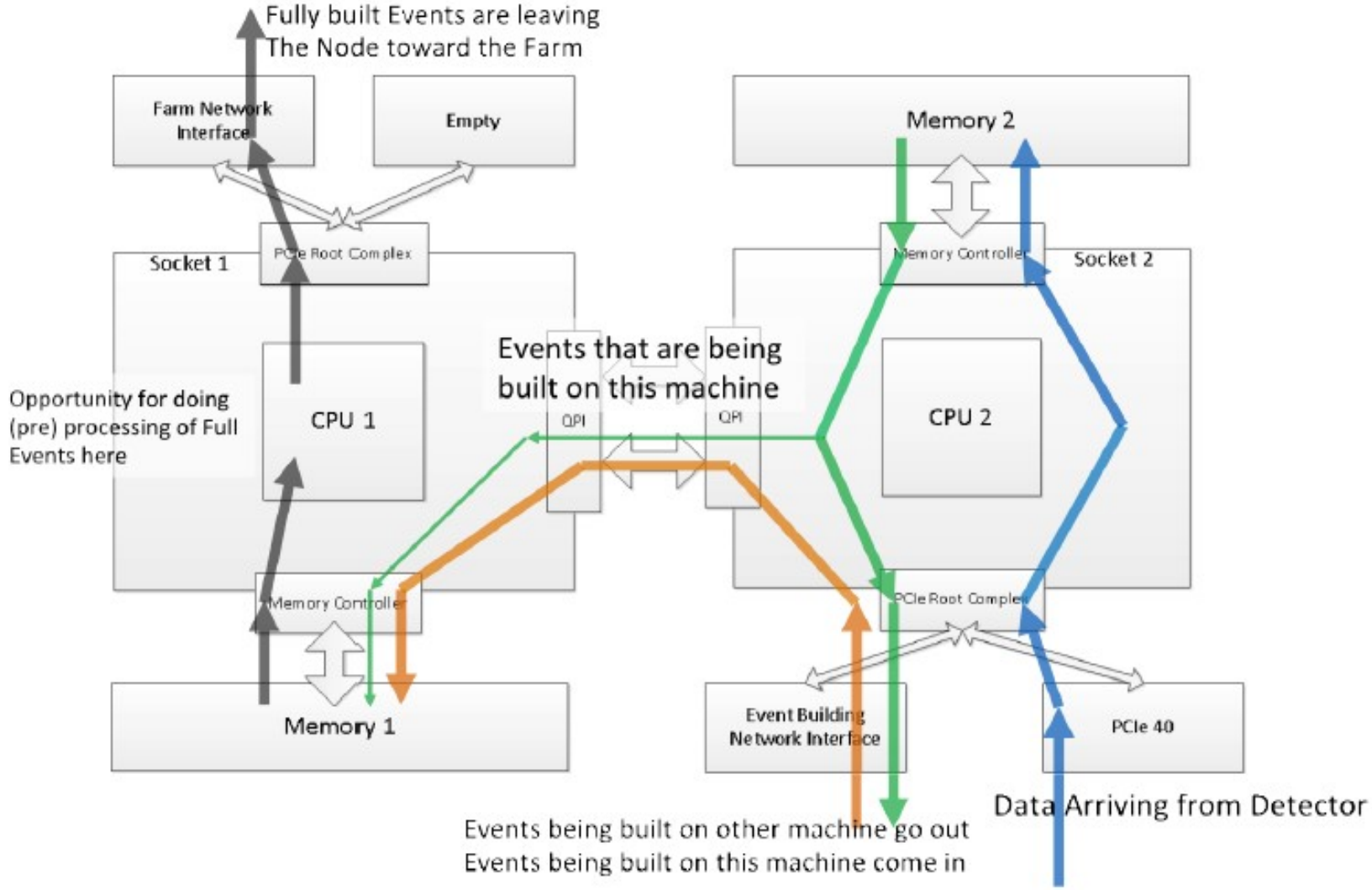


Figure 3.10: Data-flow in the event-builder server

LHCb PCIe based readout system

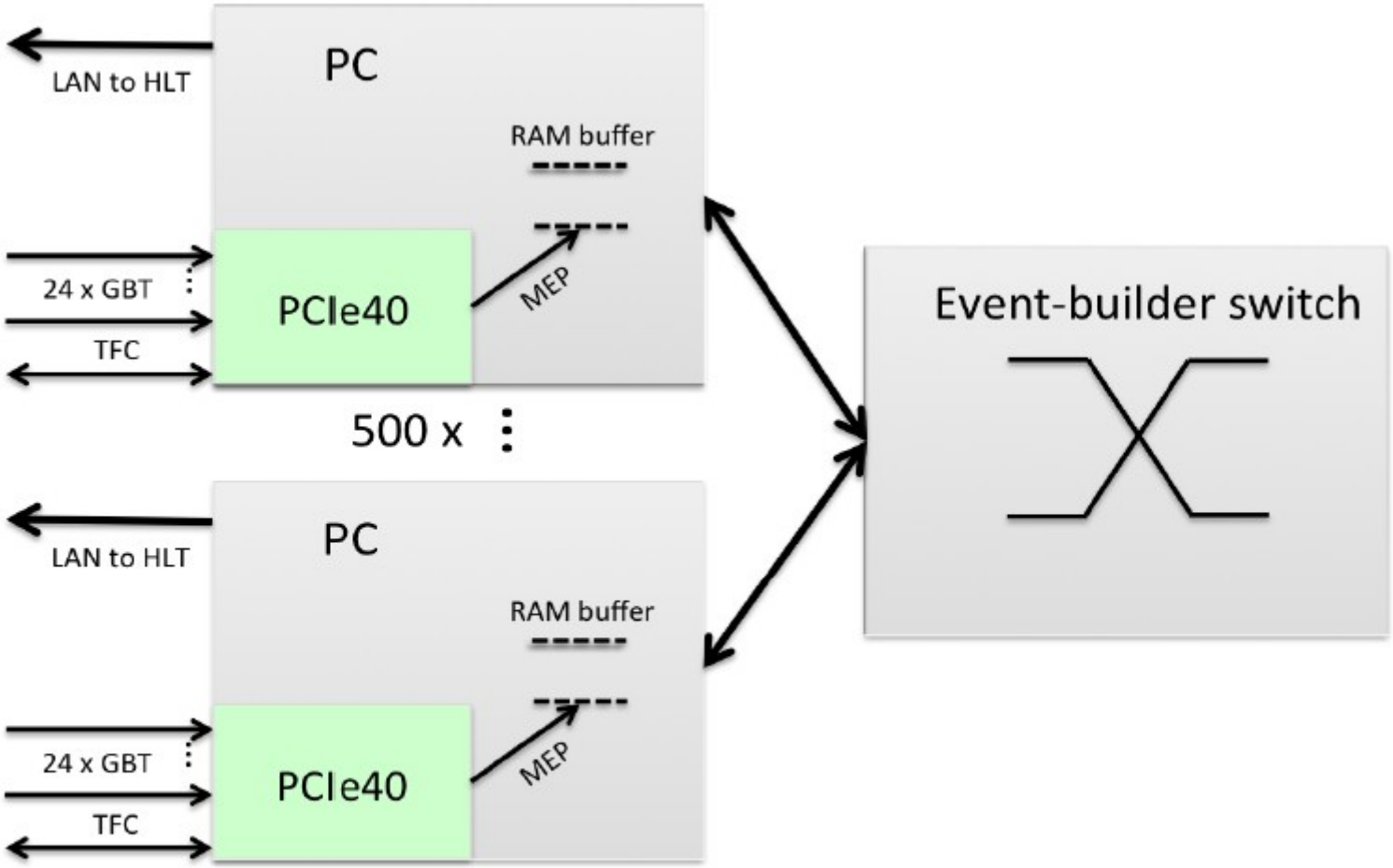
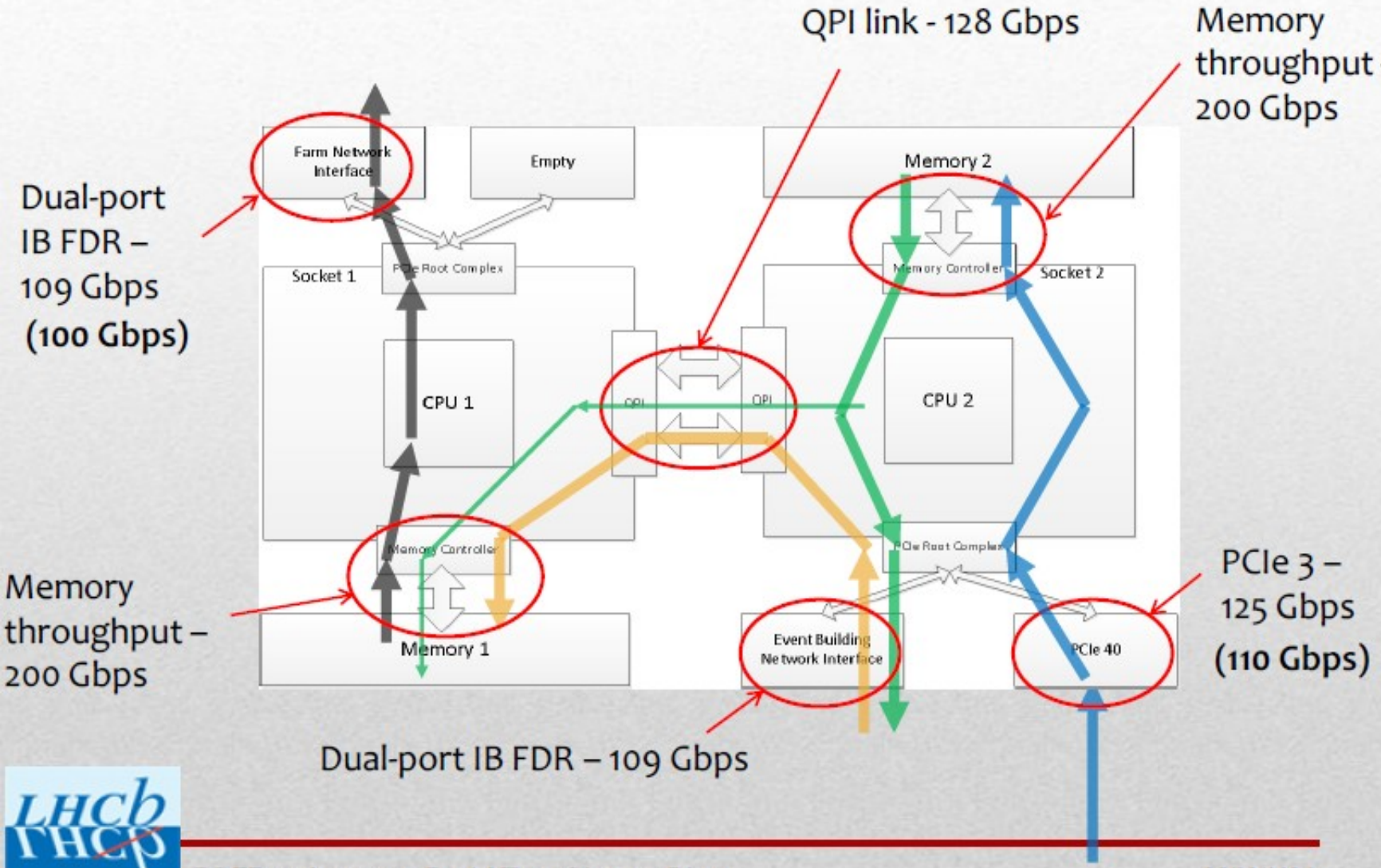
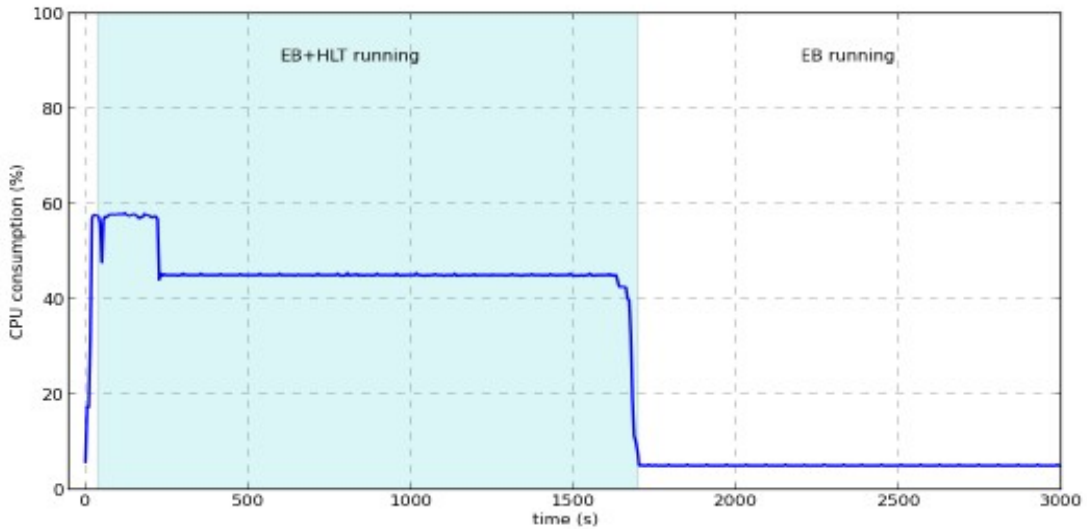


Figure 3.2: The PCIe based readout system. The PCIe40 readout boards are directly connected to the event-builder PCs through 16-lane PCIe edge-connector.

Detailed PC bandwidth

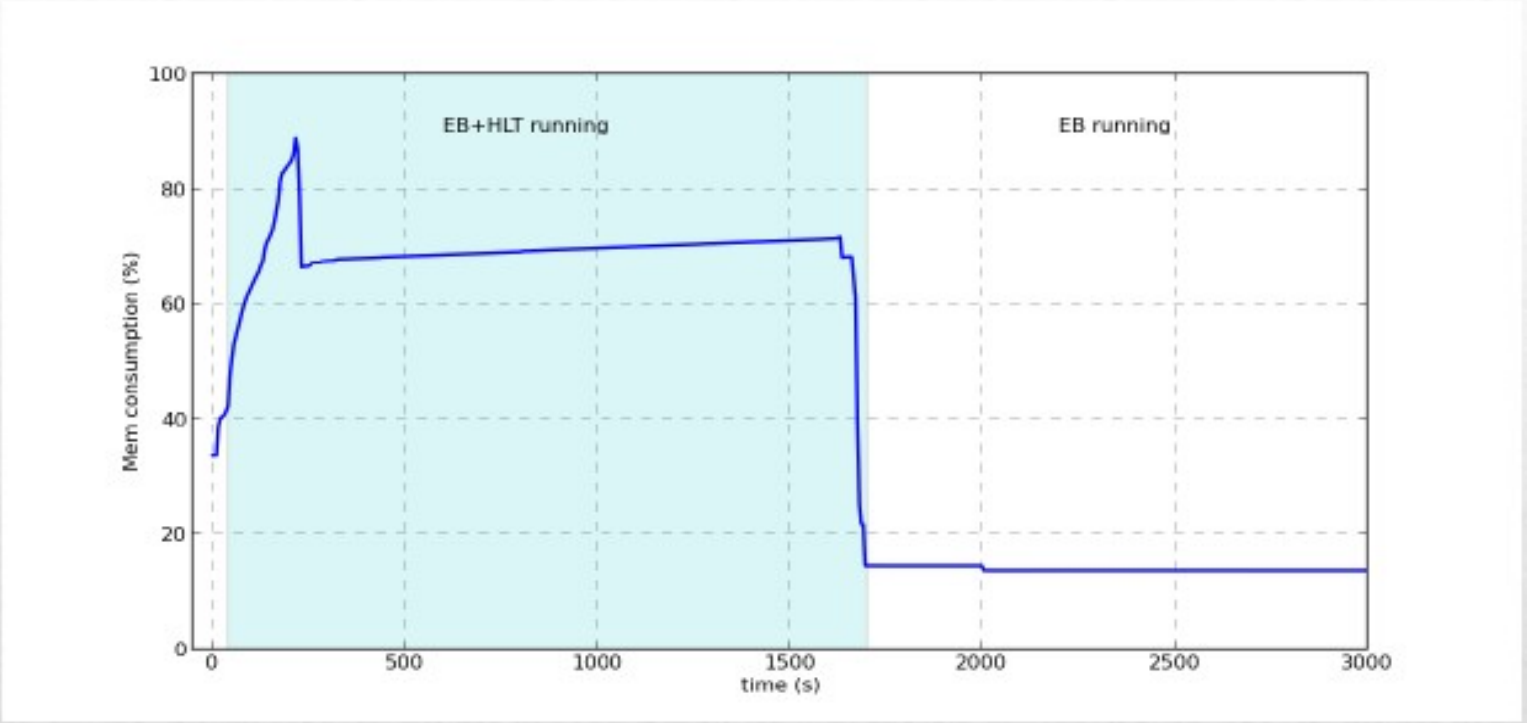


EB + trigger performance – CPU util



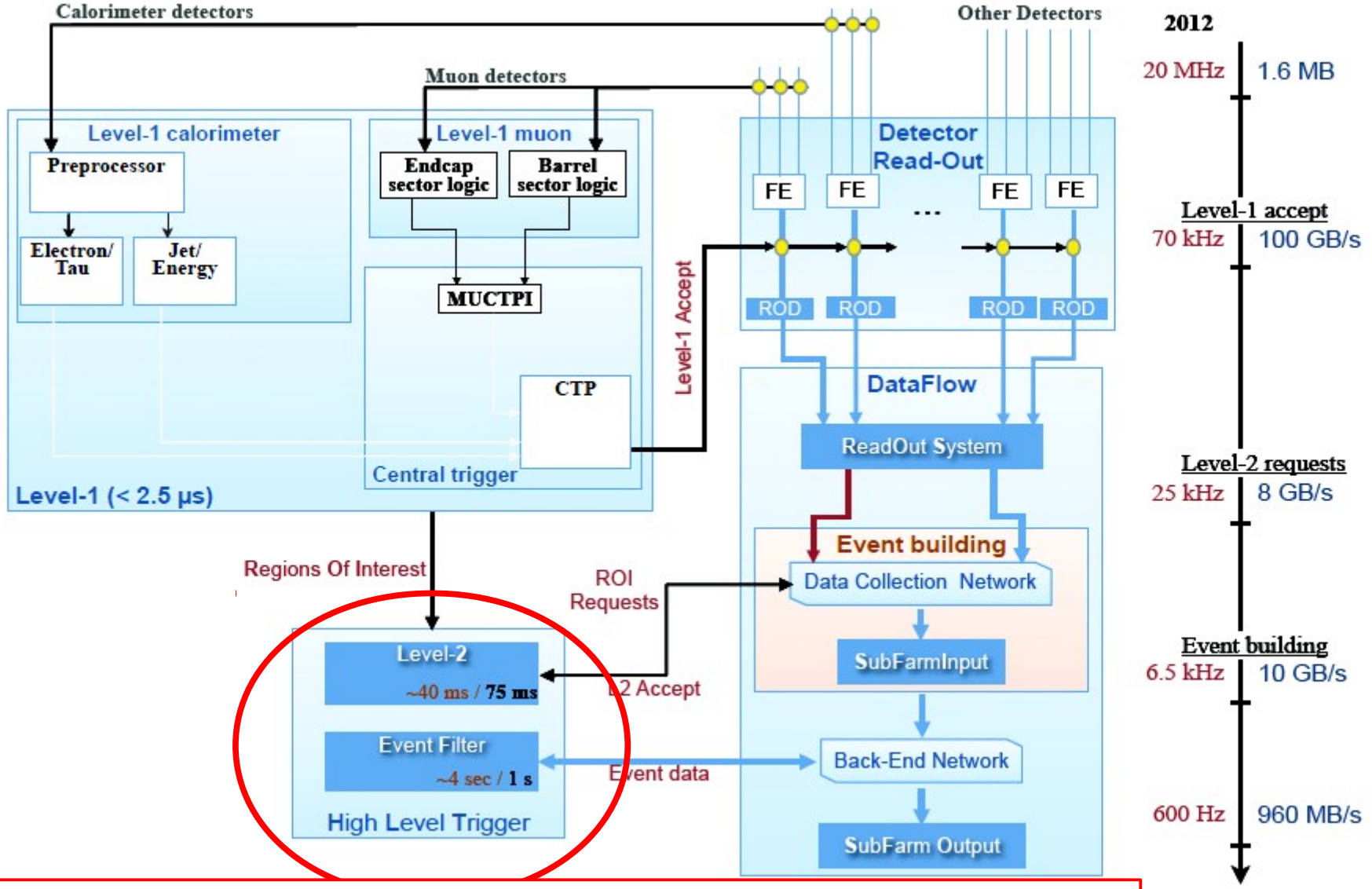
- Aggregated CPU utilization of EB application and trigger is around 46%
- Event Builder execution requires around 6 logical cores
- We are executing up to 18 instances of our Online reconstruction software (Moore)

EB + trigger performance – MEM util



- Event Builder memory utilization is currently 6 GiB
  - Input (generator) and output (filtering) buffers + application

ATLAS Current trigger



HLT = Level-2 + Event Filter (EF) farms, 2300 PC. Usually one instance of the software is run per processor core (CPU)

ATLAS: HLT in Phase 1 and Phase 2

PHASE 1

From 100 kHz from Level-1 to 1 Khz for recording

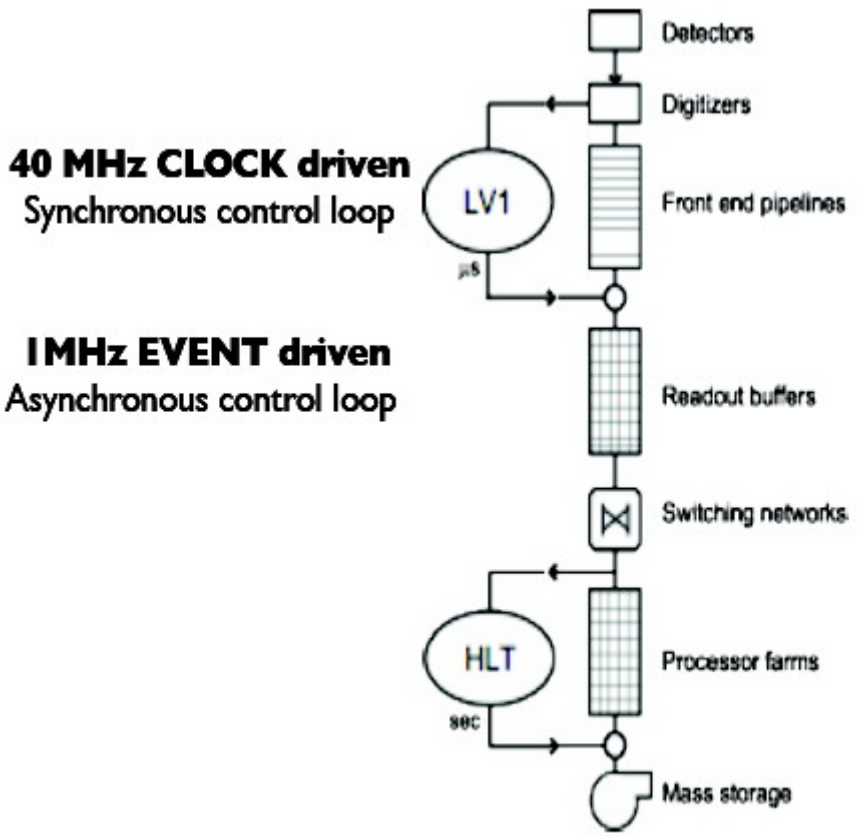
- No change to the overall dataflow architecture.
- Full calorimetry information available
- Increased use of tracks thanks to FTK
- Adopt offline processing algorithms
- New framework to support concurrent execution at multiple levels

PHASE 2

From 200 kHz from Level-1 to 5-10 kHz for recording.

- Redesign of HLT code:
  - whole event reconstruction can be performed by the HLT
  - offline-quality reconstruction for trigger objects
- Exploit new computing hardware --> significant changes to the HLT framework to support execution of algorithms in parallel threads, and to the algorithms themselves to allow deeper parallelisation.

CMS: HLT in Phase 2



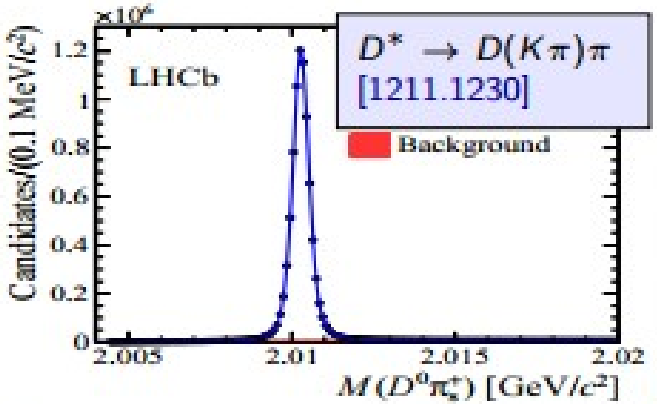
The DAQ and the HLT will be upgraded for up to 1 KHz into HLT and 10 kHz out to maintain current rejection factor

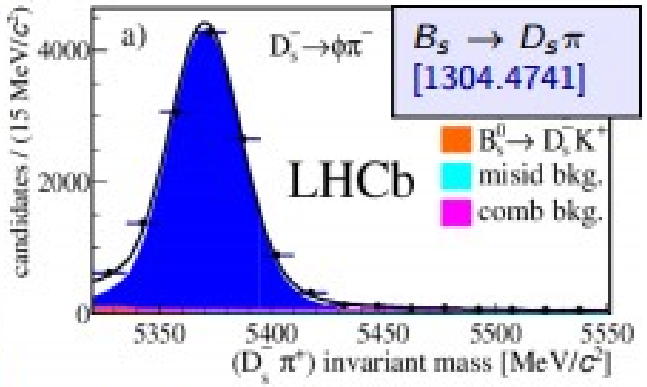
"Moore's Law" (CPUs, networks, storage) over 10 years suggests that "normal technology improvements" will handle this, including offline.

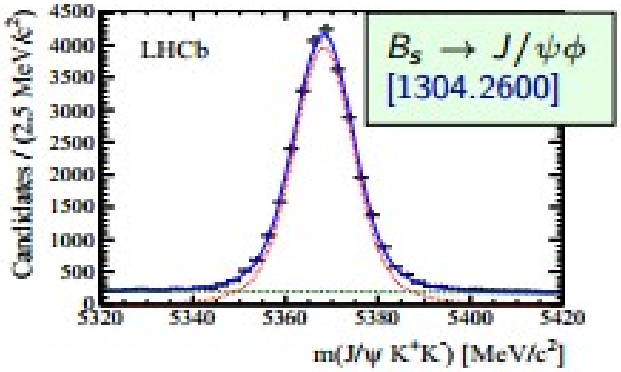
Trigger performance for selected Physics channels

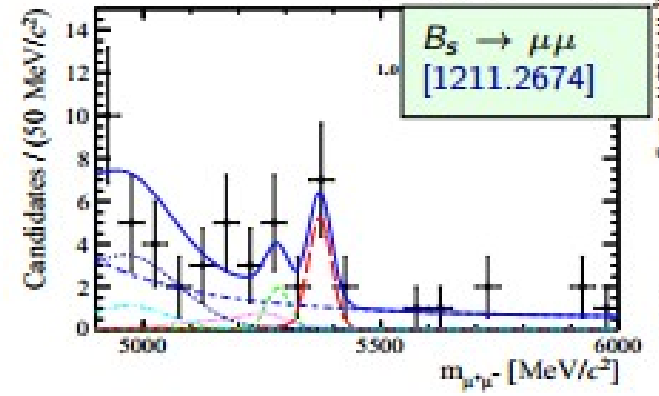
	Hadronic		Muonic	Radiative
mode	$D \to hhh$	$B \to hh$	$B^+ \to J/\psi K^+$	$B \to K^*\gamma$
ϵ(L0) [%]	27	62	93	85
ϵ(Hlt|L0) [%]	42	85	92	67
ϵ(tot) [%]	11	52	84	57

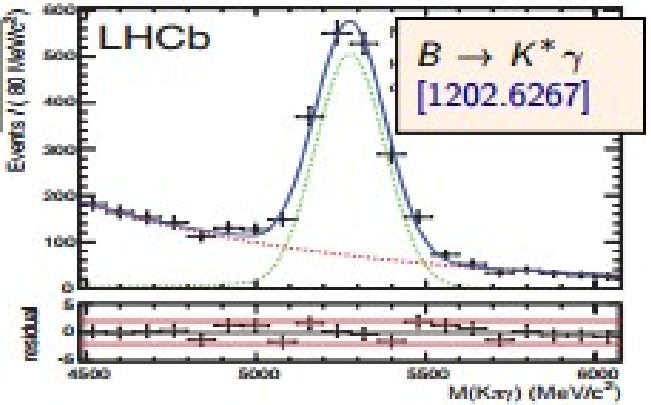
Details in [JINST 8 (2013) P04022]











Offline selected

The evolution of Network Interconnects